Introduction to Bayesian inference for astronomy

Tom Loredo
Cornell Center for Astrophysics and Planetary Science
http://www.astro.cornell.edu/staff/loredo/

CASt Summer School — 1–5 June 2021

The weather forecaster

Joint Frequencies of
Actual & Predicted Weather

	Actual		
Prediction	Rain	Sun	
Rain	**1/4**	**1/2**	3/4
Sun	**0**	**1/4**	1/4
	1/4	3/4	

Forecaster is right only 50% of the time

Observer notes a prediction of ‘Sun’ *every day* would be right 75% of the time, and applies for the forecaster’s job

Should the observer get the job?

	Actual	
Prediction	Rain	Sun
Rain	1/4	1/2
Sun	0	1/4

Forecaster: You'll never be in an unpredicted rain

Observer: You'll be in an unpredicted rain 1 day out of 4

Bayesian viewpoint

The value of an inference lies in its usefulness in the individual case

Long run performance is not an adequate criterion for assessing the usefulness of inferences

When long run performance is deemed important, it needs to be separately evaluated

Entry points for literature comparing Bayesian and frequentist approaches:

- Jaynes (1976): Confidence Intervals vs Bayesian Intervals (article # 32)
- Loredo (1992): The promise of Bayesian inference for astrophysics; also at BIPS
- Loredo (2013): Bayesian astrostatistics: A backward look to the future

Ten Great Ideas about Chance (2017)

Semi-technical survey by a leading statsitician/mathematician (Diaconis) and a leading philosopher of science (Skyrms). "This is a history book, a probability book, and a philosophy book."

"To anyone with an interest in probability or statistics, this is a book you must read. . . . [It] is far-ranging and can be read at many levels, from the novice to the expert. It is also thoroughly engaging." —David M. Bressoud, UMAP Journal

Scientific method

> *Science is more than a body of knowledge; it is a way of thinking.*
> *The method of science, as stodgy and grumpy as it may seem,*
> *is far more important than the findings of science.*
> —Carl Sagan

Scientists *argue!*

Argument $\equiv$ Collection of statements comprising an act of reasoning from *premises* to a *conclusion*

A key goal of science: Explain or predict *quantitative measurements* (data!)

Data analysis: Constructing and appraising arguments that reason from data to interesting scientific conclusions (explanations, predictions)

The role of data

Data do not speak for themselves!

> *"No body of data tells us all we need to know about its own analysis."*
>
> — John Tukey, *EDA*

We don't just *tabulate* data, we *analyze* data

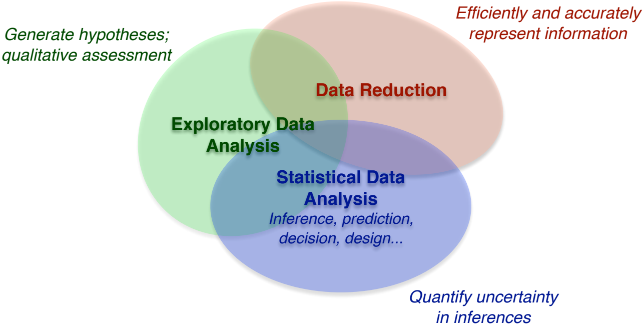
We gather data so they may speak for or against existing hypotheses, and guide the formation of new hypotheses

A key role of data in science is to be among the premises in scientific arguments

Data analysis

Building & Appraising Arguments Using Data

Modes of Data Analysis


Efficiently and accurately represent information
Generate hypotheses; qualitative assessment
Data Reduction
Exploratory Data Analysis
Statistical Data Analysis
Inference, prediction, decision, design...
Quantify uncertainty in inferences

Inference: Learning about the data generating process (population, signals. . .) from observed data—just one of several interacting modes of analyzing data

Fundamental principle

“The most fundamental principle of the statistical paradigm,
its starting point,
is that variation may be described by probability.”

Fundamental principle

~~*“The most fundamental principle of the statistical paradigm,*~~
~~*its starting point,*~~
~~*is that variation may be described by probability.”*~~

Fundamental principle

The most fundamental principle of the statistical paradigm,
its starting point,
is that ***uncertainty*** *may be described by probability.*

Fundamental principle

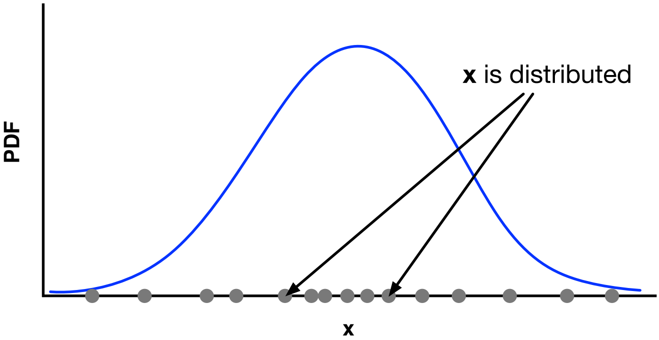
The most fundamental principle of the statistical paradigm,
its starting point,
is that ***uncertainty*** *may be described by probability.*

An important corollary is that, in some settings
—most notably, for ***IID replications****,*
and for ***exchangeable sequences****—*
expected variation and individual-case probability
are intimately linked.

Interpreting PDFs

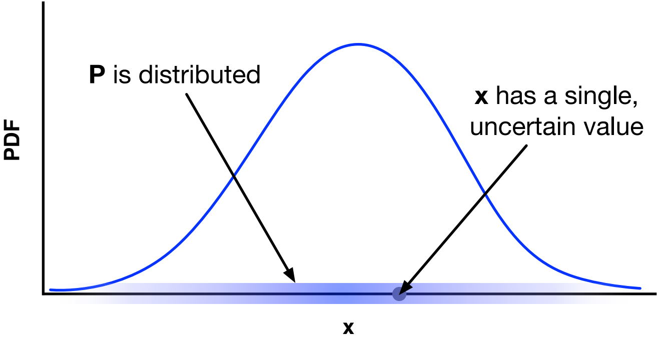
Frequentist

Probabilities are always (limiting) rates/proportions/frequencies that *quantify variability* in a sequence of trials. $p(x)$ describes how the *values of x* would be distributed among infinitely many trials:



Bayesian

Probability *quantifies uncertainty* in an inductive inference. $p(x)$ describes how *probability* is distributed over the possible values x might have taken in the single case before us:



Probability & frequency in IID settings

Frequency from probability

Bernoulli's (weak) law of large numbers: In repeated IID trials, given $P(\text{success}|\ldots) = \alpha$, predict

$$\frac{n_{\text{success}}}{N_{\text{total}}} \to \alpha \quad \text{as} \quad N_{\text{total}} \to \infty$$

If $P(\text{success}|\ldots)$ does not change from sample to sample, it may be interpreted as a relative frequency

Probability from frequency

Bayes's "An Essay Towards Solving a Problem in the Doctrine of Chances" → First use of Bayes's theorem:

Probability for success in next trial of IID sequence:

$$\mathsf{E}(\alpha) \to \frac{n_{\text{success}}}{N_{\text{total}}} \quad \text{as} \quad N_{\text{total}} \to \infty$$

If $P(\text{success}|\ldots)$ does not change from sample to sample, it may be estimated using relative frequency data

Twiddle notation for the normal distribution

$$\mathrm{Norm}(x;\mu,\sigma) \equiv \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{\sigma^2}\right]$$

Frequentist

$$p(\overset{\textit{random}}{x}\;;\;\overset{\textit{fixed but unknown}}{\mu,\sigma}) = \mathrm{Norm}(x,\mu,\sigma)$$

$$x \sim \mathcal{N}(\mu,\sigma^2)$$

"*x is distributed* as normal with mean…"

"random" = "varies unpredictably in repeated trials"

Bayesian

$$p(\overset{\textit{random}}{x}\;|\;\overset{\textit{random or known}}{\mu,\sigma}) = \mathrm{Norm}(x,\mu,\sigma)$$

$$x \sim \mathcal{N}(\mu,\sigma^2)$$

"*The probability for x is distributed* as normal with mean…"

"random" = "uncertain in the case at hand"

Bayesian statistical inference

- Bayesian inference uses probability theory to *quantify the strength of data-based arguments* (i.e., a more abstract view than restricting PT to describe variability in repeated "random" experiments)
- A different approach to *all* statistical inference problems (i.e., not just another method in the list: BLUE, linear regression, least squares/χ^2 minimization, maximum likelihood, ANOVA, survival analysis, LDA classification . . .)
- Focuses on *deriving consequences of modeling assumptions* rather than *devising and calibrating procedures*

Bayesian data analysis (BDA): Using Bayesian ideas across various data analysis tasks—not just inference, but also prediction, decision, design, EDA, data reduction. . .

Frequentist vs. Bayesian statements

"The data D_{obs} support conclusion C . . . "

Frequentist assessment

"C was selected with a procedure that's right 95% of the time over a set $\{D_{hyp}\}$ that includes D_{obs}."

Probabilities are properties of *procedures*, not of particular results.

Bayesian assessment

"The strength of the chain of reasoning from the model and D_{obs} to C is 0.95, on a scale where 1= certainty."

Probabilities are associated with *specific, observed data*. Long-run performance must be separately evaluated (and is typically good by frequentist criteria).

Agenda

1. **Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$**

   Confidence intervals vs. credible intervals

2. **Probability theory as generalized logic**

3. **Probability theory for data analysis: Three theorems**

4. **Inference with parametric models**

   Parameter Estimation

   Model Uncertainty

5. **Quick-looks**

   Curve fitting & least squares

   Measurement error & hierarchical/graphical models

   Bayesian computation menu

Agenda

1. **Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$**
   - Confidence intervals vs. credible intervals
2. **Probability theory as generalized logic**
3. **Probability theory for data analysis: Three theorems**
4. **Inference with parametric models**
   - Parameter Estimation
   - Model Uncertainty
5. **Quick-looks**
   - Curve fitting & least squares
   - Measurement error & hierarchical/graphical models
   - Bayesian computation menu

A Simple (?) confidence region

Problem

Estimate the location (mean, μ) of a Gaussian distribution from a set of N IID samples $D = \{x_i\}$. Report a region summarizing the uncertainty.

Here assume std dev'n σ is *known*; we are uncertain only about μ

Model

The *sampling distribution* for *any* set $\{x_i\}$ is

$$\begin{aligned} p(\{x_i\}|\mu) &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}; \qquad \sigma = 1 \\ &\propto e^{-\chi^2(\mu)/2} \end{aligned}$$

This gives the *likelihood function*, $\mathcal{L}(\mu)$ if we set $\{x_i\}$ to the *observed values*

Classes of variables—the two spaces

- μ is the unknown we seek to estimate—the *parameter*. The *parameter space* is the space of possible values of μ—here the real line (perhaps bounded). *Hypothesis space* is a more general term.
- A particular set of N data values $D = \{x_i\}$ is a *sample*. The *sample space* is the N-dimensional space of possible samples.

Standard inferences

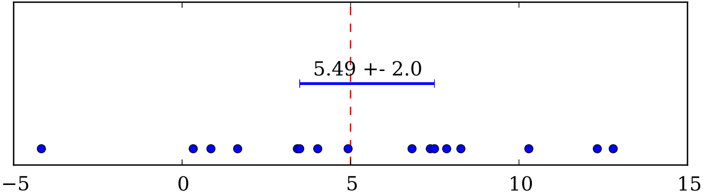
Let $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$.

- "Standard error" (rms error) is $\sigma/\sqrt{N}$
- "1σ" interval: $\bar{x} \pm \sigma/\sqrt{N}$ with conf. level $\mathrm{CL} = 68.3\%$
- "2σ" interval: $\bar{x} \pm 2\sigma/\sqrt{N}$ with $\mathrm{CL} = 95.4\%$

Some simulated data

Take $\mu = 5$ and $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$.

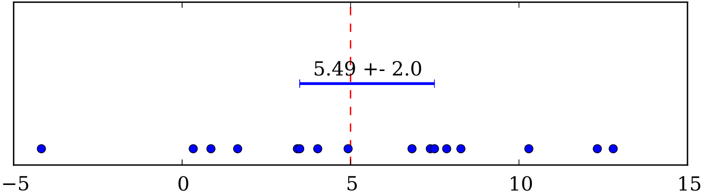
What is the CL associated with this interval?



Some simulated data

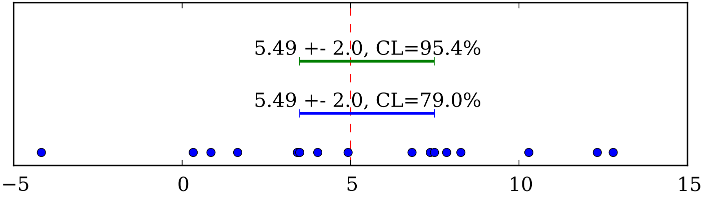
Take $\mu = 5$ and $\sigma = 4$ and $N = 16$, so $\sigma/\sqrt{N} = 1$.

What is the CL associated with this interval?



The confidence level for this interval is 79.0%.

Two intervals



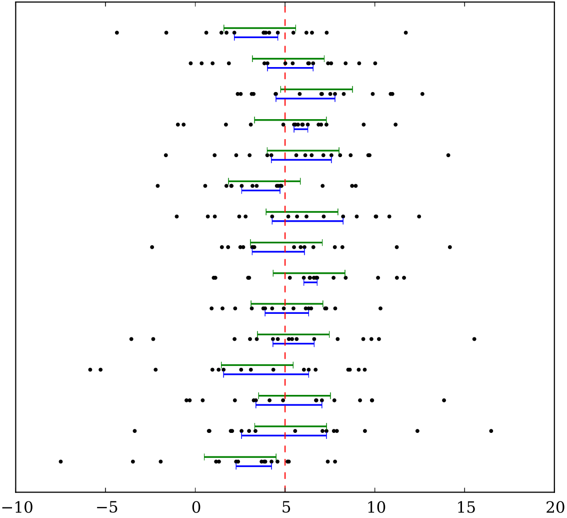
- Green interval: $\bar{x} \pm 2\sigma/\sqrt{N}$
- Blue interval: Let $x_{(k)} \equiv k$'th order statistic
  Report $[x_{(6)}, x_{(11)}]$ (i.e., leave out 5 outermost each side)

The point

> *The (frequentist) confidence level is a property of the* ***procedure****, not of the particular interval reported for a given dataset*

Performance of intervals

Intervals for 15 datasets



Confidence interval for a normal mean

Suppose we have a sample of $N = 5$ values x_i,

$$x_i \sim N(\mu, 1)$$

We want to estimate μ, including some *quantification of uncertainty* in the estimate: an interval *with a probability attached*

Frequentist approaches: method of moments, BLUE, least-squares/χ^2, maximum likelihood

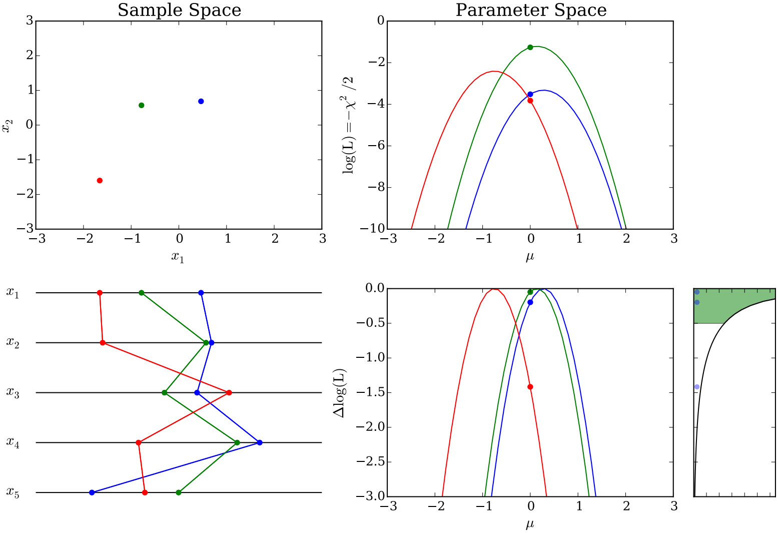
Focus on likelihood (equivalent to χ^2 here); this is closest to Bayes:

$$\begin{aligned} \mathcal{L}(\mu) &\equiv p(\{x_i\}|\mu) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/2\sigma^2}; \qquad \sigma = 1 \\ &\propto e^{-\chi^2(\mu)/2} \end{aligned}$$

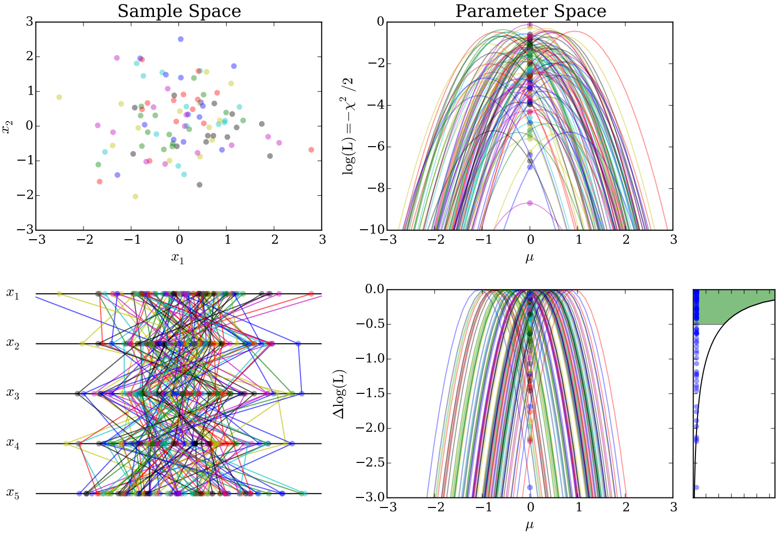
Estimate μ from maximum likelihood (minimum χ^2)
Define an interval and its coverage frequency from the $\mathcal{L}(\mu)$ curve

Construct an interval procedure for known μ

Likelihoods for 3 simulated data sets, $\mu = 0$



Likelihoods for 100 simulated data sets, $\mu = 0$



Careful! This is for $\mu = 0$, but μ will be unknown.
Luckily, the $\Delta\log(\mathcal{L})$ dist'n is independent of μ.

Credible interval for a normal mean

Recall the likelihood, $\mathcal{L}(\mu) \equiv p(D_{\text{obs}}|\mu)$, is a probability for the observed data, but *not* for the parameter μ (wrong PDF units)

Convert likelihood to a probability distribution over μ via *Bayes's theorem* (changes units from D to μ):

$$\begin{aligned} p(A, B) &= p(A)p(B|A) \\ &= p(B)p(A|B) \\ \rightarrow p(A|B) &= p(A)\frac{p(B|A)}{p(B)}, \quad \text{Bayes's th.} \end{aligned}$$

$$\Rightarrow p(\mu|D_{\text{obs}}) \quad \propto \quad \pi(\mu)\mathcal{L}(\mu)$$

$p(\mu|D_{\text{obs}})$ is called the *posterior probability distribution*

This requires a prior probability density, $\pi(\mu)$, often taken to be constant over the allowed region if there is no significant information available (or sometimes constant w.r.t. some reparameterization motivated by a symmetry in the problem)

Gaussian problem posterior distribution

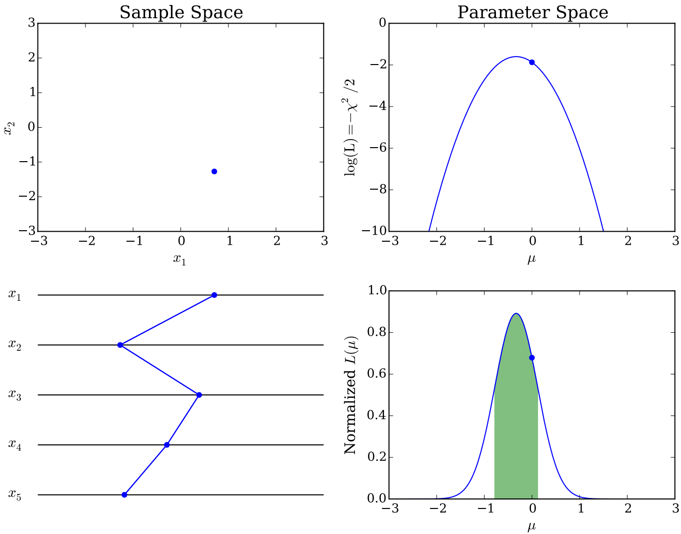
For the Gaussian example, a bit of algebra ("complete the square") gives:

$$\begin{aligned}
\mathcal{L}(\mu) &\propto \prod_i \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
&\propto \exp\left[-\frac{1}{2}\sum_i \frac{(x_i-\mu)^2}{\sigma^2}\right] \\
&\propto \exp\left[-\frac{(\mu-\bar{x})^2}{2(\sigma/\sqrt{N})^2}\right]
\end{aligned}$$

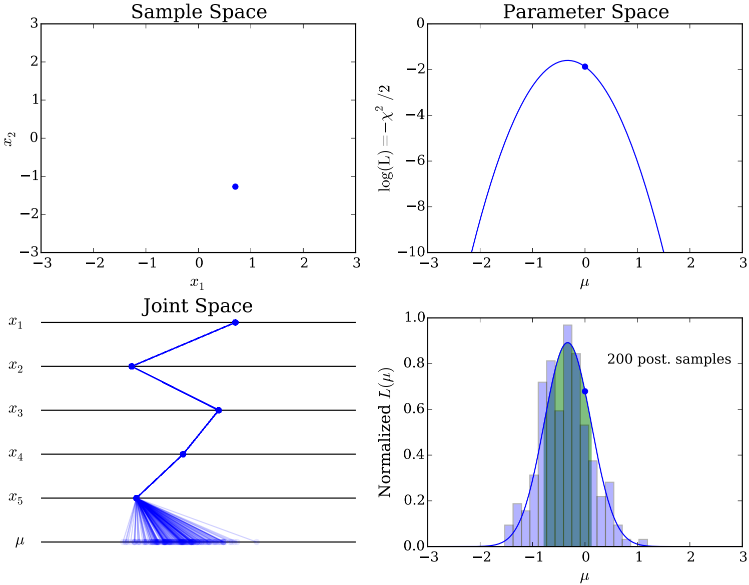
The likelihood is Gaussian in μ.
Flat prior → posterior density for μ is $\mathcal{N}(\bar{x}, \sigma^2/N)$.

Bayesian credible region

Normalize the likelihood for the observed sample; report the region that includes 68.3% of the normalized likelihood:



Posterior sampling: Credible region via Monte Carlo (MCMC, ABC)



Posterior summaries

- Posterior mean is $\langle\mu\rangle \equiv \int d\mu\, \mu\, p(\mu|D_{\rm obs}) = \bar{x}$
- Posterior mode is $\hat{\mu} = \bar{x}$
- Posterior std dev'n is $\sigma/\sqrt{N}$
- $\bar{x} \pm \sigma/\sqrt{N}$ is a 68.3% *credible region*:

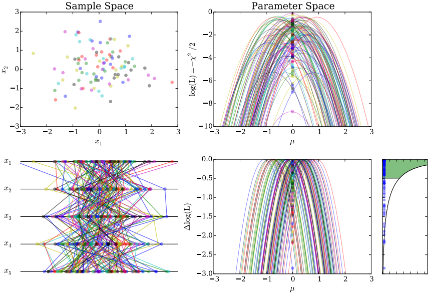
$$\int_{\bar{x}-\sigma/\sqrt{N}}^{\bar{x}+\sigma/\sqrt{N}} d\mu\ p(\mu|D_{\rm obs}) \approx 0.683$$

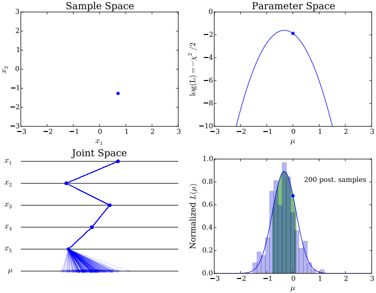
- $\bar{x} \pm 2\sigma/\sqrt{N}$ is a 95.4% credible region

The credible regions above are *highest posterior density* credible regions (HPD regions). These are the smallest regions with a specified probability content.

These reproduce familiar frequentist results, but this is a *coincidence* due to special properties of Gaussians.

Confidence vs. credible regions





When the approaches differ

Both approaches report $\mu \in [\bar{x} - \sigma/\sqrt{N}, \bar{x} + \sigma/\sqrt{N}]$, and assign 68.3% to this interval (with different meanings)

This matching is a *coincidence*!

When might results differ? ($\mathcal{F}$ = frequentist, $\mathcal{B}$ = Bayes)

- If $\mathcal{F}$ procedure doesn't use likelihood directly
- If $\mathcal{F}$ procedure properties depend on params (e.g., nonlinear models; need to find pivotal quantities)
- If likelihood shape varies strongly between datasets (conditional inference, ancillary statistics, recognizable subsets)
- If there are extra uninteresting parameters (nuisance parameters; adjusted profile likelihood, conditional inference)
- If $\mathcal{B}$ uses important prior information

Also, for a different task—comparison of parametric models—the approaches are *qualitatively* different (significance tests & info criteria vs. Bayes factors)

Bayesian and Frequentist inference

Brad Efron, ASA President (2005)

> The 250-year debate between Bayesians and frequentists is unusual among philosophical arguments in actually having *important practical consequences*. . . . The physicists I talked with were really bothered by our 250 year old Bayesian-frequentist argument. Basically there's only one way of doing physics but there seems to be at least two ways to do statistics, and *they don't always give the same answers*. . . .
>
> Broadly speaking, Bayesian statistics dominated 19th Century statistical practice while the 20th Century was more frequentist. What's going to happen in the 21st Century?. . . I strongly suspect that statistics is in for a burst of new theory and methodology, and that this burst will feature a combination of Bayesian and frequentist reasoning. . . .

Roderick Little, ASA President's Address (2005)

Pragmatists might argue that good statisticians can get sensible answers under Bayes or frequentist paradigms; indeed maybe two philosophies are better than one, since they provide more tools for the statistician's toolkit. . . . I am discomforted by this "inferential schizophrenia." Since the Bayesian (B) and frequentist (F) philosophies *can differ even on simple problems*, at some point decisions seem needed as to which is right. I believe our credibility as statisticians is undermined when we cannot agree on the fundamentals of our subject. . . .

An assessment of strengths and weaknesses of the frequentist and Bayes systems of inference suggests that *calibrated Bayes*. . . captures the strengths of both approaches and provides a roadmap for future advances.

[Calibrated Bayes = Bayesian inference within a specified space of models + frequentist-based model checking; Andrew Gelman et al. use "Bayesian data analysis" similarly]

(see TL's arXiv:1208.3035 for discussion/references)

Agenda

1. **Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$**
   - Confidence intervals vs. credible intervals
2. **Probability theory as generalized logic**
3. **Probability theory for data analysis: Three theorems**
4. **Inference with parametric models**
   - Parameter Estimation
   - Model Uncertainty
5. **Quick-looks**
   - Curve fitting & least squares
   - Measurement error & hierarchical/graphical models
   - Bayesian computation menu

Logic—some essentials

"Logic can be defined as *the analysis and appraisal of arguments*"
—Gensler, *Intro to Logic*

Build arguments with propositions and logical operators/connectives:

- *Propositions:* Statements that may be true or false

$\mathcal{P}$:	Universe can be modeled with ΛCDM
A :	$\Omega_{\rm tot} \in [0.9, 1.1]$
B :	Ω_Λ is not 0
$\overline{B}$:	"not B," i.e., $\Omega_\Lambda = 0$

- *Connectives:*

$A \wedge B$:	A and B are *both* true
$A \vee B$:	A or B is true, or both are

Arguments

Argument: Assertion that an *hypothesized conclusion*, H, follows from *premises*, $\mathcal{P} = \{A, B, C, \ldots\}$ (take "," = "and")

Notation:

$H|\mathcal{P}$: Premises $\mathcal{P}$ imply H
H may be deduced from $\mathcal{P}$
H follows from $\mathcal{P}$
H is true given that $\mathcal{P}$ is true

Arguments are (compound) propositions.

Central role of arguments → special terminology for true/false:

- A true argument is *valid*
- A false argument is *invalid* or *fallacious*

Valid vs. sound arguments

Content vs. form

- An argument is *factually correct* iff all of its *premises are true* (it has "good content").
- An argument is *valid* iff its conclusion *follows from* its premises (it has "good form").
- An argument is *sound* iff it is both *factually correct and valid* (it has good form and content).

Deductive logic (and probability theory) addresses *validity*.

We want to make *sound* arguments. There is no formal approach for addressing factual correctness $\rightarrow$ there is always a subjective element to an argument.

Deductive and inductive inference

Deduction—Syllogism as prototype

Premise 1: A implies H
Premise 2: A is true
Deduction: $\therefore H$ is true
$H|\mathcal{P}$ is valid

Induction—Analogy as prototype

Premise 1: A, B, C, D, E all share properties x, y, z
Premise 2: F has properties x, y
Induction: F has property z
"F has z"$|\mathcal{P}$ is not strictly valid, but may still be rational (likely, plausible, probable); some such arguments are stronger than others

Boolean algebra (and/or/not over $\{0, 1\}$) quantifies deduction

Bayesian probability theory (and/or/not over $[0, 1]$) generalizes this to quantify the strength of inductive arguments

Representing induction with $[0,1]$ calculus

$P(H|\mathcal{P}) \equiv$ strength of argument $H|\mathcal{P}$

$$\begin{aligned} P &= 1 \rightarrow \text{Argument is } \textit{deductively valid} \\ &= 0 \rightarrow \text{Premises imply } \overline{H} \\ &\in (0,1) \rightarrow \text{Degree of deducibility} \end{aligned}$$

Mathematical model for induction

$$\text{`AND' (product rule):} \quad P(A \wedge B|\mathcal{P}) = P(A|\mathcal{P})\,P(B|A \wedge \mathcal{P}) = P(B|\mathcal{P})\,P(A|B \wedge \mathcal{P})$$

$$\text{`OR' (sum rule):} \quad P(A \vee B|\mathcal{P}) = P(A|\mathcal{P}) + P(B|\mathcal{P}) - P(A \wedge B|\mathcal{P})$$

$$\text{`NOT':} \quad P(\overline{A}|\mathcal{P}) = 1 - P(A|\mathcal{P})$$

Pierre Simon Laplace (1819)

Probability theory is nothing but *common sense reduced to calculation*.

James Clerk Maxwell (1850)

They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic, but the actual science of *Logic is conversant at present only with things either certain, impossible, or entirely doubtful*, none of which (fortunately) we have to reason on. Therefore *the true logic of this world is the calculus of Probabilities*, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

Harold Jeffreys (1931)

If we like there is no harm in saying that a probability expresses a degree of reasonable belief. . . . 'Degree of confirmation' has been used by Carnap, and possibly avoids some confusion. But whatever verbal expression we use to try to convey the primitive idea, this expression cannot amount to a definition. *Essentially the notion can only be described by reference to instances where it is used.* It is intended to express *a kind of relation between data and consequence* that habitually arises in science and in everyday life, and the reader should be able to recognize the relation from examples of the circumstances when it arises.

Agenda

1. **Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$**
   Confidence intervals vs. credible intervals
2. **Probability theory as generalized logic**
3. **Probability theory for data analysis: Three theorems**
4. **Inference with parametric models**
   Parameter Estimation
   Model Uncertainty
5. **Quick-looks**
   Curve fitting & least squares
   Measurement error & hierarchical/graphical models
   Bayesian computation menu

The Bayesian Recipe

Assess hypotheses by calculating their probabilities $p(H_i | \ldots)$ conditional on known and/or presumed information (including observed data) using the rules of probability theory.

Probability Theory Axioms

$\mathcal{C} \equiv$ context, initial set of premises

$$\text{`OR' (sum rule):} \quad P(H_1 \vee H_2|\mathcal{C}) = P(H_1|\mathcal{C}) + P(H_2|\mathcal{C}) - P(H_1, H_2|\mathcal{C})$$

$$\begin{aligned} \text{`AND' (product rule):} \quad P(H_i, D_{\text{obs}}|\mathcal{C}) &= P(H_i|\mathcal{C})\, P(D_{\text{obs}}|H_i, \mathcal{C}) \\ &= P(D_{\text{obs}}|\mathcal{C})\, P(H_i|D_{\text{obs}}, \mathcal{C}) \end{aligned}$$

$$\text{`NOT':} \quad P(\overline{H_i}|\mathcal{C}) = 1 - P(H_i|\mathcal{C})$$

Three Important Theorems

Bayes's Theorem (BT)

Consider $P(H_i, D_{\rm obs}|\mathcal{C})$ using the product rule:

$$\begin{aligned} P(H_i, D_{\rm obs}|\mathcal{C}) &= P(H_i|\mathcal{C})\, P(D_{\rm obs}|H_i, \mathcal{C}) \\ &= P(D_{\rm obs}|\mathcal{C})\, P(H_i|D_{\rm obs}, \mathcal{C}) \end{aligned}$$

Solve for the *posterior probability* (expands the premises!):

$$P(H_i|D_{\rm obs}, \mathcal{C}) = P(H_i|\mathcal{C})\, \frac{P(D_{\rm obs}|H_i, \mathcal{C})}{P(D_{\rm obs}|\mathcal{C})}$$

Theorem holds for any propositions, but for hypotheses & data the factors have names:

posterior $\propto$ *prior* $\times$ *likelihood*

norm. const. $P(D_{\rm obs}|\mathcal{C}) =$ *prior predictive*

Law of Total Probability (LTP)

Consider exclusive, exhaustive $\{B_i\}$ ($\mathcal{C}$ asserts one of them must be true),

$$\begin{aligned}\sum_i P(A, B_i|\mathcal{C}) &= \sum_i P(B_i|A, \mathcal{C})P(A|\mathcal{C}) = P(A|\mathcal{C}) \\ &= \sum_i P(B_i|\mathcal{C})P(A|B_i, \mathcal{C})\end{aligned}$$

If we do not see how to get $P(A|\mathcal{C})$ directly, we can find a set $\{B_i\}$ and use it as a "basis"—*extend the conversation*:

$$P(A|\mathcal{C}) = \sum_i P(B_i|\mathcal{P})P(A|B_i, \mathcal{C})$$

If our problem already has B_i in it, we can use LTP to get $P(A|\mathcal{P})$ from the joint probabilities—*marginalization*:

$$P(A|\mathcal{C}) = \sum_i P(A, B_i|\mathcal{C})$$

Example: Take $\mathcal{P} = \mathcal{C}$, $A = D_{\rm obs}$, $B_i = H_i$; then

$$\begin{aligned} P(D_{\rm obs}|\mathcal{C}) &= \sum_i P(D_{\rm obs}, H_i|\mathcal{C}) \\ &= \sum_i P(H_i|\mathcal{C}) P(D_{\rm obs}|H_i, \mathcal{C}) \end{aligned}$$

prior predictive for $D_{\rm obs}$ = Average likelihood for H_i
(a.k.a. *marginal likelihood*)

Normalization

For *exclusive, exhaustive* H_i,

$$\sum_i P(H_i|\cdots) = 1$$

Tabular/diagrammatic Bayesian inference

Simplest case: *Binary classification*

- 2 hypotheses: $\{C, \overline{C}\}$
- 2 possible data values: $\{-, +\}$

Concrete example: You test positive $(+)$ for a medical condition. Do you have the condition (C) or not $(\overline{C})$?

- Prior: Prevalence of the condition in your population is 0.1%
- Likelihood:
  - Test is 80% accurate if you have the condition: $P(+|C, \mathcal{C}) = 0.8$ ("sensitivity")
  - Test is 95% accurate if you are healthy: $P(-|\overline{C}, \mathcal{C}) = 0.95$ ("specificity," $1 - p(\text{false } +)$)

Numbers roughly correspond to mammography screening for breast cancer in asymptomatic women

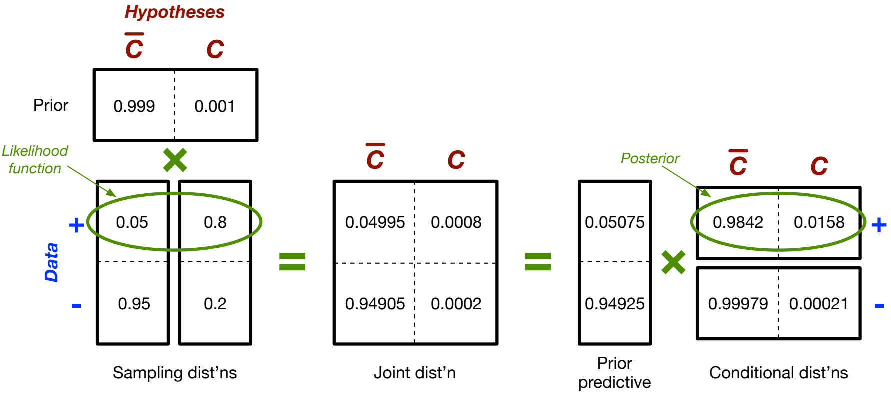
Tabular calculation

Hypothesis H_i	Prior $\pi_i \equiv p(H_i)$	Likelihood $\mathcal{L}_i \equiv p(+\|H_i)$	Joint $\pi_i \times \mathcal{L}_i$	Posterior $p(H_i\|+)$
$\overline{C}$	0.999	0.05	0.04995	0.9842
C	0.001	0.8	0.0008	0.0158
Sums:	1.0	**NA**	0.05075 $= p(+)$	1.0

Inference as manipulation of the joint distribution

Bayes's theorem in terms of the *joint distribution*:

$$P(H_i|\mathcal{C}) \times P(D_{\rm obs}|H_i,\mathcal{C}) = P(H_i, D_{\rm obs}|\mathcal{C}) = P(H_i|D_{\rm obs},\mathcal{C}) \times P(D_{\rm obs}|\mathcal{C})$$



Hypotheses

	$\bar{C}$	C
Prior	0.999	0.001

×

Likelihood function

Data	$\bar{C}$	C
+	0.05	0.8
-	0.95	0.2

Sampling dist'ns

=

$\bar{C}$	C
0.04995	0.0008
0.94905	0.0002

Joint dist'n

=

0.05075
0.94925

Prior predictive

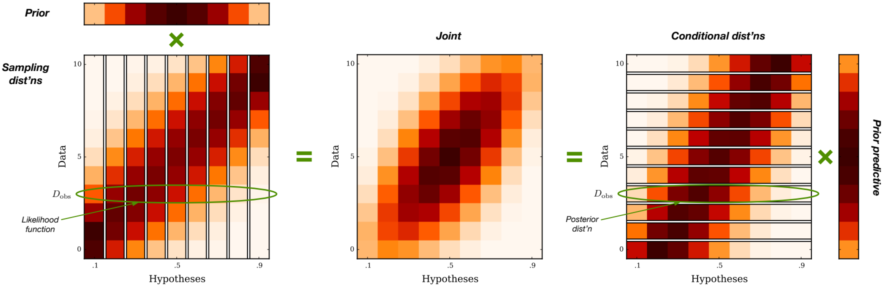
×

Posterior

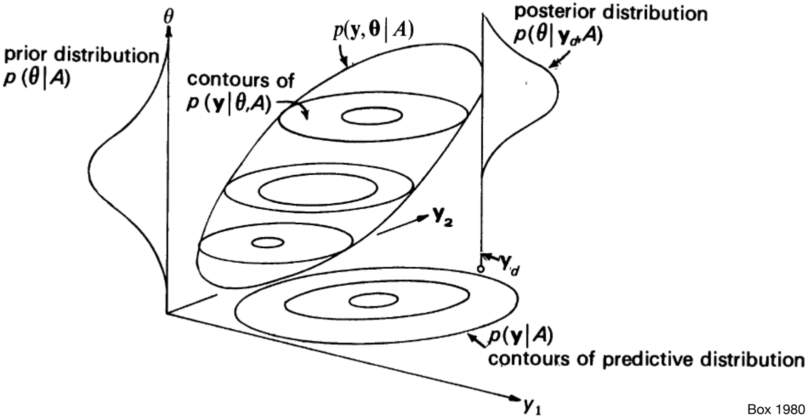
$\bar{C}$	C	
0.9842	0.0158	+
0.99979	0.00021	-

Conditional dist'ns

Larger discrete case: 9 hypotheses, 11 possible data values



Continuous data, parameter spaces



θ
prior distribution
$p(\theta|A)$
$p(\mathbf{y}, \boldsymbol{\theta}|A)$
contours of
$p(\mathbf{y}|\theta, A)$
posterior distribution
$p(\theta|\mathbf{y}_d, A)$
$\mathbf{y}_2$
$\mathbf{y}_d$
$p(\mathbf{y}|A)$
contours of predictive distribution
y_1

Box 1980

Components of Bayes's theorem for a problem with a 1-D parameter space (θ) and a 2-D sample space (**y**), with observed data $\mathbf{y}_\mathrm{d}$, and modeling assumptions *A*

Recap of Key Ideas

Probability as generalized logic

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

Bayes's theorem

$$p(\text{Hypothesis} \mid \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})$$

Data *change* the support for a hypothesis $\propto$ ability of hypothesis to *predict* the data

Law of total probability

$$p(\text{Hypothes}\underline{\textbf{es}} \mid \text{Data}) = \sum p(\text{Hypothes}\underline{\textbf{is}} \mid \text{Data})$$

The support for a *compound/composite* hypothesis must account for all the ways it could be true

Agenda

1. **Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$**
   - Confidence intervals vs. credible intervals

2. **Probability theory as generalized logic**

3. **Probability theory for data analysis: Three theorems**

4. **Inference with parametric models**
   - Parameter Estimation
   - Model Uncertainty

5. **Quick-looks**
   - Curve fitting & least squares
   - Measurement error & hierarchical/graphical models
   - Bayesian computation menu

Inference With Parametric Models

Models M_i ($i = 1$ to N), each with parameters θ_i, each imply a *sampling dist'n* (conditional predictive dist'n for possible data):

$$p(D|\theta_i, M_i)$$

The θ_i dependence when we fix attention on the *observed* data is the *likelihood function*:

$$\mathcal{L}_i(\theta_i) \equiv p(D_{\text{obs}}|\theta_i, M_i)$$

We may be uncertain about i (model uncertainty) or θ_i (parameter uncertainty).

Henceforth we will only consider the actually observed data, so we drop the cumbersome subscript: $D = D_{obs}$.

Classes of Problems

Single-model inference

Premise = choice of single model (specific i)
Parameter estimation: What can we say about θ_i or $f(\theta_i)$?
Prediction: What can we say about future data D'?

Multi-model inference

Premise = $\{M_i\}$
Model comparison/choice: What can we say about i?
Model averaging:

- *Systematic error*: $\theta_i = \{\phi, \eta_i\}$; ϕ is common to all
  What can we say about ϕ w/o committing to one model?
- *Prediction*: What can we say about future D', accounting for model uncertainty?

Model checking

Premise = $M_1 \vee$ "all" alternatives
Is M_1 adequate? (predictive tests, calibration, robustness)

Parameter Estimation

Problem statement

$\mathcal{C}$ = Model M with parameters θ (+ any add'l info)

H_i = statements about θ; e.g. "$\theta \in [2.5, 3.5]$," or "$\theta > 0$"

Probability for any such statement can be found using a *probability density function* (PDF) for θ:

$$\begin{aligned} P(\theta \in [\theta, \theta + d\theta] | \cdots) &= f(\theta) d\theta \\ &= p(\theta | \cdots) d\theta \end{aligned}$$

Posterior probability density

$$p(\theta | D, M) = \frac{p(\theta | M)\ \mathcal{L}(\theta)}{\int d\theta\ p(\theta | M)\ \mathcal{L}(\theta)}$$

Summaries of posterior

- "Best fit" values:
  - *Mode*, $\hat{\theta}$, maximizes $p(\theta|D,M)$
  - *Posterior mean*, $\langle\theta\rangle = \int d\theta\,\theta\,p(\theta|D,M)$
- Uncertainties:
  - *Credible region* Δ of probability C:
    $C = P(\theta \in \Delta|D,M) = \int_\Delta d\theta\,p(\theta|D,M)$
    Highest Posterior Density (HPD) region has $p(\theta|D,M)$ higher inside than outside
  - Posterior standard deviation, variance, covariances
- Marginal distributions
  - Interesting parameters ϕ, nuisance parameters η
  - *Marginal dist'n* for ϕ: $p(\phi|D,M) = \int d\eta\,p(\phi,\eta|D,M)$

Estimating a Normal Mean

Problem specification

Model: $d_i = \mu + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, σ is known → $I = (\sigma, M)$.

Parameter space: μ; seek $p(\mu|D, \sigma, M)$

Likelihood

$$\begin{aligned}
\mathcal{L}(\mu) &\equiv p(D|\mu, \sigma, M) \\
&= \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-(d_i-\mu)^2/2\sigma^2}; \qquad \sigma = 1 \\
&\propto \exp\left(-\frac{N(\mu - \overline{d})^2}{2\sigma^2}\right)
\end{aligned}$$

Likelihood function is a Gaussian function at $\overline{d}$, width $w = \sigma/\sqrt{N}$

Informative Conjugate Prior

Use a normal prior, $\mu \sim N(\mu_0, w_0^2)$

Conjugate because the posterior turns out also to be normal

$w_0 \to \infty$ is the "uninformative" flat prior limit; posterior remains normal and proper (normalizable)

Posterior

Normal $N(\tilde{\mu}, \tilde{w}^2)$, but mean, std. deviation *"shrink"* towards prior.

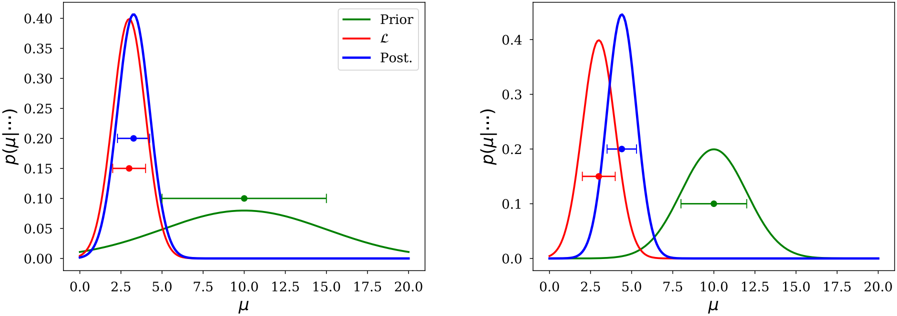
Define $B = \frac{w^2}{w^2+w_0^2}$, so $B < 1$ and $B = 0$ when w_0 is large. Then

$$\begin{aligned} \widetilde{\mu} &= \overline{d} + B \cdot (\mu_0 - \overline{d}) \\ \widetilde{w} &= w \cdot \sqrt{1 - B} \end{aligned}$$

"Principle of stable estimation" — The prior affects estimates only when data are not informative relative to prior

Conjugate normal examples:

- Data have $\overline{d} = 3$, $\sigma/\sqrt{N} = 1$
- Priors at $\mu_0 = 10$, with $w = \{5, 2\}$



Supplement:

- Binomial example
  - Bernoulli trials: Bernoulli process & binomial sampling dist'ns
  - Beta-binomial conjugate model
- Normal example
  - Analytical details for normal example
  - Sufficiency; sample mean and variance as sufficient statistics
  - Handling σ uncertainty by marginalizing over σ; Student's t distribution

Nuisance Parameters and Marginalization

To model most data, we need to introduce parameters besides those of ultimate interest: *nuisance parameters*.

Example

> We have data from measuring a rate $r = s + b$ that is a sum of an interesting signal s and a background b.
>
> We have additional data just about b.
>
> What do the data tell us about s?

Marginal posterior distribution

To summarize implications for s, accounting for b uncertainty, *marginalize*:

$$\begin{aligned} p(s|D,M) &= \int db\; p(s,b|D,M) \\ &\propto p(s|M) \int db\; p(b|s,M)\, \mathcal{L}(s,b) \\ &= p(s|M)\mathcal{L}_m(s) \end{aligned}$$

with $\mathcal{L}_m(s)$ the *marginal likelihood function for s*:

$$\mathcal{L}_m(s) \equiv \int db\; p(b|s)\, \mathcal{L}(s,b)$$

Marginalization vs. Profiling

For insight: Suppose the prior is broad compared to the likelihood → for a fixed s, we can accurately estimate b with max likelihood $\hat{b}_s$, with small uncertainty δb_s.

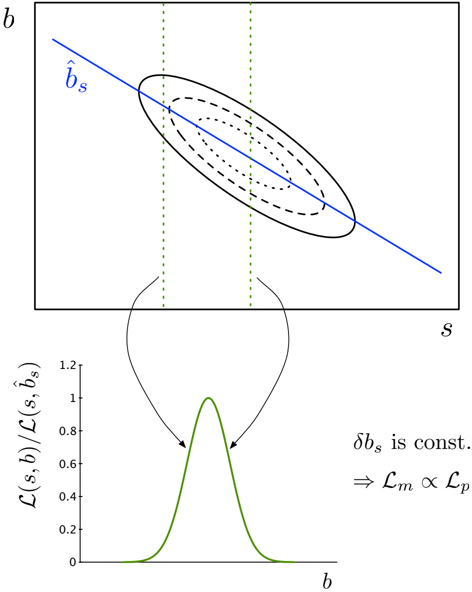
$$
\begin{aligned}
\mathcal{L}_m(s) &\equiv \int db\; p(b|s)\,\mathcal{L}(s,b) \\
&\approx p(\hat{b}_s|s)\,\mathcal{L}(s,\overset{\text{best } b \text{ given } s}{\hat{b}_s})\;\underset{b \text{ uncertainty given } s}{\delta b_s}
\end{aligned}
$$

Profile likelihood $\mathcal{L}_p(s) \equiv \mathcal{L}(s,\hat{b}_s)$ gets weighted by a *parameter space volume factor*

E.g., Gaussians: $\hat{s} = \hat{r} - \hat{b}, \quad \sigma_s^2 = \sigma_r^2 + \sigma_b^2$

Background *subtraction* is a special case of background *marginalization*.

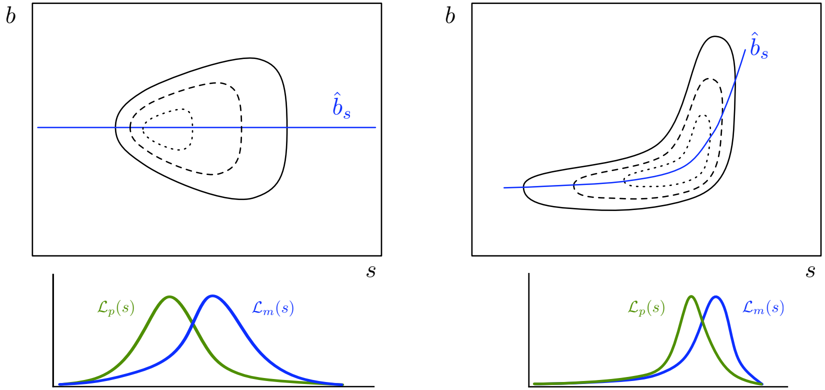
Bivariate normals: $\mathcal{L}_m \propto \mathcal{L}_p$



δb_s is const. vs. s

$\Rightarrow \mathcal{L}_m \propto \mathcal{L}_p$

Flared/skewed/bannana-shaped: $\mathcal{L}_m$ and $\mathcal{L}_p$ *differ*



General result: For a linear (in params) model sampled with Gaussian noise, and flat priors, $\mathcal{L}_m \propto \mathcal{L}_p$.
Otherwise, they will likely *differ*.

In *"measurement error problems"* the difference can be dramatic

The On/Off Problem for Poisson counting data

Basic problem

- Look off-source; unknown background rate b
  Count $N_{\rm off}$ photons in interval $T_{\rm off}$
- Look on-source; rate is $r = s + b$ with unknown signal s
  Count $N_{\rm on}$ photons in interval $T_{\rm on}$
- Infer s

Conventional solution

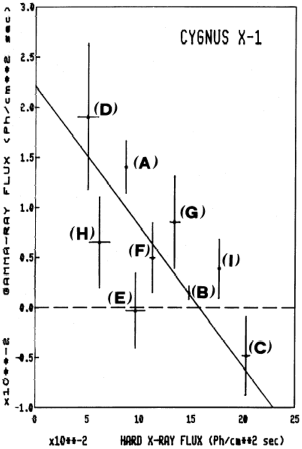
$$\begin{aligned}
\hat{b} &= N_{\rm off}/T_{\rm off}; & \sigma_b &= \sqrt{N_{\rm off}}/T_{\rm off} \\
\hat{r} &= N_{\rm on}/T_{\rm on}; & \sigma_r &= \sqrt{N_{\rm on}}/T_{\rm on} \\
\hat{s} &= \hat{r} - \hat{b}; & \sigma_s &= \sqrt{\sigma_r^2 + \sigma_b^2}
\end{aligned}$$

But $\hat{s}$ can be *negative*!

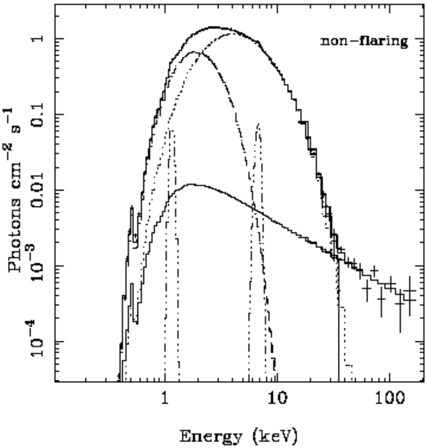
Examples

Spectra of X-Ray Sources

Bassani et al. 1989

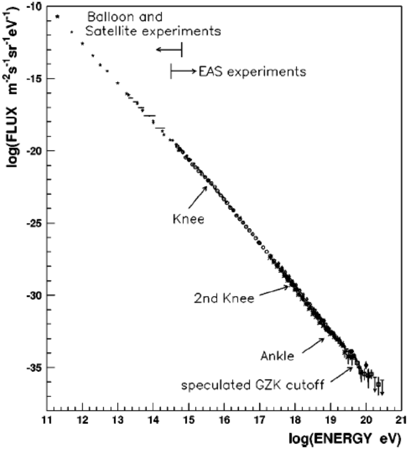


Di Salvo et al. 2001

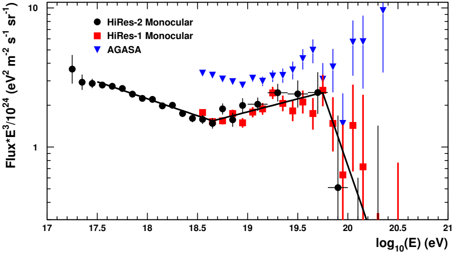


Spectrum of Ultrahigh-Energy Cosmic Rays

Nagano & Watson 2000



HiRes Team 2007



N is Never Large

Sample sizes are never large. If N is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once N is 'large enough,' you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). N is never enough because if it were 'enough' you'd already be on to the next problem for which you need more data.

— Andrew Gelman (blog entry, 31 July 2005)

N is Never Large

Sample sizes are never large. If N is too small to get a sufficiently-precise estimate, you need to get more data (or make more assumptions). But once N is 'large enough,' you can start subdividing the data to learn more (for example, in a public opinion poll, once you have a good estimate for the entire country, you can estimate among men and women, northerners and southerners, different age groups, etc etc). N is never enough because if it were 'enough' you'd already be on to the next problem for which you need more data.

Similarly, you never have quite enough money. But that's another story.

— Andrew Gelman (blog entry, 31 July 2005)

Bayesian Solution to On/Off Problem

The likelihood function is a product of separate Poisson distributions for the off-source and on-source data:

$$\mathcal{L}(s,b) = \frac{(bT_{\rm off})^{N_{\rm off}}}{N_{\rm off}!} e^{-bT_{\rm off}} \times \frac{[(s+b)T_{\rm on}]^{N_{\rm on}}}{N_{\rm on}!} e^{-(s+b)T_{\rm on}}$$

Adopting flat priors for (s, b), the joint posterior is

$$p(s,b|N_{\rm on},N_{\rm off},\mathcal{C}) \quad \propto \quad (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})}$$

Recall if $b = 0$, the (normalized) posterior distribution is a gamma distribution,

$$p(s,b=0|N_{\rm on},N_{\rm off},\mathcal{C}) = \frac{T_{\rm on}(sT_{\rm on})^{N_{\rm on}}}{N_{\rm on}!} e^{-sT_{\rm on}}$$

Now marginalize over b;

$$\begin{aligned} p(s|N_{\rm on}, N_{\rm off}, \mathcal{C}) &= \int db\; p(s, b \mid N_{\rm on}, \mathcal{C}) \\ &\propto \int db\; (s+b)^{N_{\rm on}} b^{N_{\rm off}} e^{-sT_{\rm on}} e^{-b(T_{\rm on}+T_{\rm off})} \end{aligned}$$

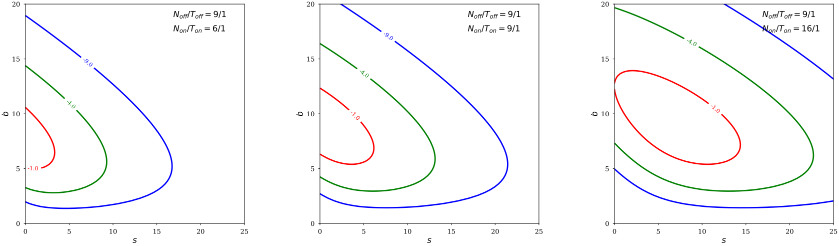
Expand $(s+b)^{N_{\rm on}}$ and do the resulting Γ integrals:

$$\begin{aligned} p(s|N_{\rm on}, N_{\rm off}, \mathcal{C}) &= \sum_{i=0}^{N_{\rm on}} C_i \frac{T_{\rm on}(sT_{\rm on})^i e^{-sT_{\rm on}}}{i!} \\ C_i &\propto \left(1 + \frac{T_{\rm off}}{T_{\rm on}}\right)^i \frac{(N_{\rm on} + N_{\rm off} - i)!}{(N_{\rm on} - i)!} \end{aligned}$$

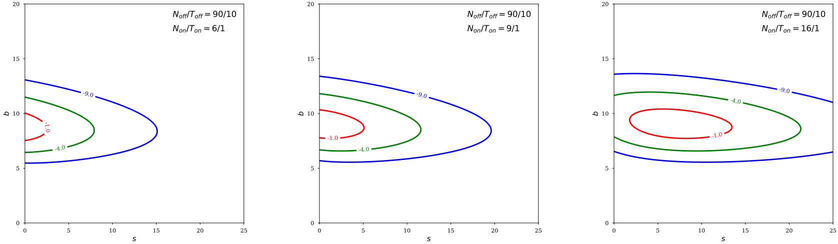
Posterior is a weighted sum of Gamma distributions, each assigning a different number of on-source counts to the source. (Evaluate via recursive algorithm or confluent hypergeometric function.)

Example on/off joint PDFs

$T_{\rm off} = 1$

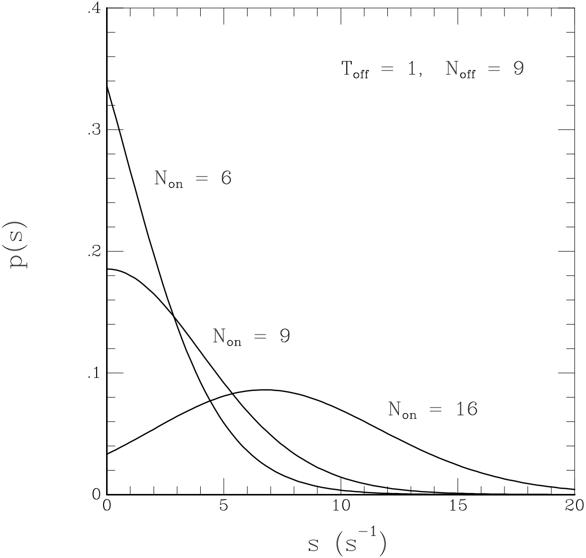


$T_{\rm off} = 10$



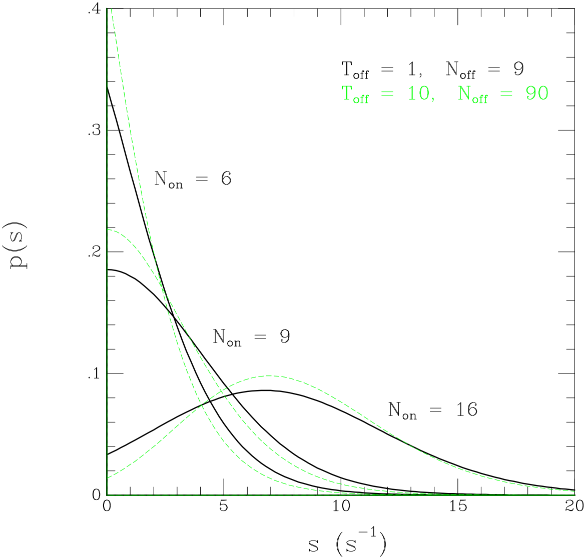
Example on/off marginal PDFs—Short integrations

$T_{\mathrm{on}} = T_{\mathrm{off}} = 1$



Example on/off marginal PDFs—Long background integrations

$T_{\rm on} = 1$, $T_{\rm off} = 10$



Supplement:

- Analytical details for Poisson dist'n inference
- Gamma-Poisson conjugate model
- Alternative (equivalent) solution to the on/off problem
- Multibin case

Many Roles for Marginalization

Eliminate nuisance parameters

$$p(\phi|D,M) = \int d\eta\; p(\phi,\eta|D,M)$$

Propagate uncertainty

Model has parameters θ; what can we infer about $F = f(\theta)$?

$$\begin{aligned} p(F|D,M) &= \int d\theta\; p(F,\theta|D,M) \;=\; \int d\theta\; p(\theta|D,M)\, p(F|\theta,M) \\ &= \int d\theta\; p(\theta|D,M)\, \delta[F - f(\theta)] \qquad [\textit{single-valued case}] \end{aligned}$$

Prediction

Given a model with parameters θ and present data D, predict future data D' (e.g., for *experimental design*):

$$p(D'|D,M) = \int d\theta\; p(D',\theta|D,M) = \int d\theta\; p(\theta|D,M)\, p(D'|\theta,M)$$

Model comparison. . .

Agenda

1. **Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$**
   - Confidence intervals vs. credible intervals
2. **Probability theory as generalized logic**
3. **Probability theory for data analysis: Three theorems**
4. **Inference with parametric models**
   - Parameter Estimation
   - Model Uncertainty
5. **Quick-looks**
   - Curve fitting & least squares
   - Measurement error & hierarchical/graphical models
   - Bayesian computation menu

Model comparison

Problem statement

$\mathcal{C} = (M_1 \vee M_2 \vee \ldots)$ — Specify a set of models.
$H_i = M_i$ — Hypothesis chooses a model.

Posterior probability for a model

$$\begin{aligned} p(M_i|D,\mathcal{C}) &= p(M_i|\mathcal{C}) \frac{p(D|M_i,\mathcal{C})}{p(D|\mathcal{C})} \\ &\propto p(M_i|\mathcal{C})\mathcal{L}(M_i) \end{aligned}$$

$$\mathcal{L}(M_i) \equiv p(D|M_i) = \int d\theta_i \, p(\theta_i|M_i) p(D|\theta_i, M_i)$$

Likelihood for model = Average likelihood for its parameters

$$\mathcal{L}(M_i) = \langle \mathcal{L}(\theta_i) \rangle$$

Varied terminology: Prior predictive = *Marginal likelihood* = Average likelihood = Global likelihood = (Weight of) Evidence for model

Odds and Bayes factors

A ratio of probabilities for two propositions using the same premises is called the *odds* favoring one over the other:

$$\begin{aligned} O_{ij} &\equiv \frac{p(M_i|D,\mathcal{C})}{p(M_j|D,\mathcal{C})} \\ &= \frac{p(M_i|\mathcal{C})}{p(M_j|\mathcal{C})} \times \frac{p(D|M_i,\mathcal{C})}{p(D|M_j,\mathcal{C})} \end{aligned}$$

The data-dependent part is called the *Bayes factor*:

$$B_{ij} \equiv \frac{p(D|M_i,\mathcal{C})}{p(D|M_j,\mathcal{C})}$$

It is a *likelihood ratio*; the BF terminology is usually reserved for cases when the likelihoods are marginal/average likelihoods for *composite hypotheses*

An Automatic Ockham's Razor

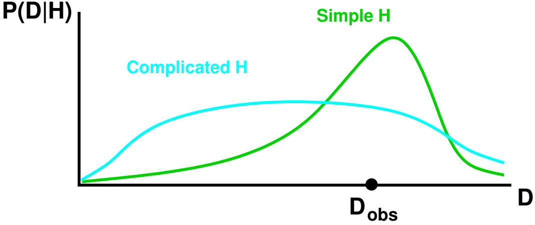
Consider *nested models*:

- Simpler model M_1 with parameters θ_1
- "Larger" rival M_2 with parameters $\theta_2 = (\theta_1, \eta)$

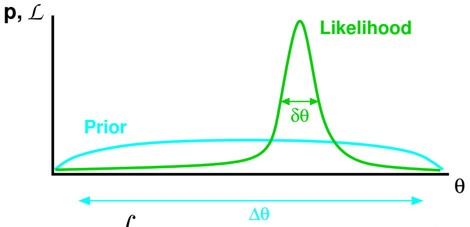
$\Rightarrow \mathcal{L}(\hat{\theta}_2) \geq \mathcal{L}(\hat{\theta}_1)$
But what about $p(D|M_i) = \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i)$?

Prior predictive distributions

Normalization implies *there must be data that favor M_1*:



The Ockham Factor

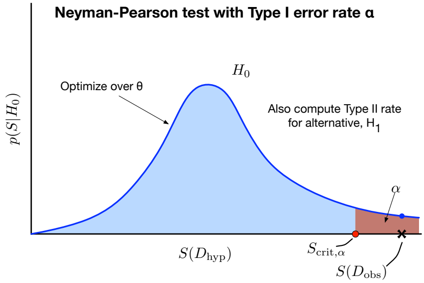


$$
\begin{aligned}
p(D|M_i) &= \int d\theta_i \; p(\theta_i|M) \; \mathcal{L}(\theta_i) \approx p(\hat{\theta}_i|M)\mathcal{L}(\hat{\theta}_i)\delta\theta_i \\
&\approx \mathcal{L}(\hat{\theta}_i)\frac{\delta\theta_i}{\Delta\theta_i} \\
&= \text{Maximum Likelihood} \times \text{Ockham Factor}
\end{aligned}
$$

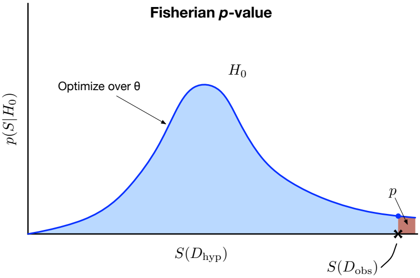
Models with more parameters often make the data more probable — *for the best fit*

Ockham factor penalizes models for "wasted" volume of parameter space

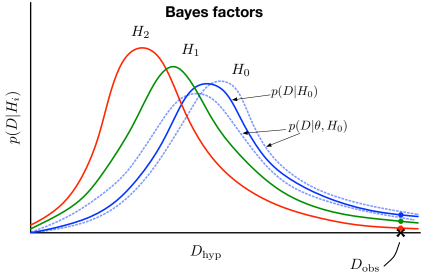
Quantifies intuition that models shouldn't require fine-tuning



Neyman-Pearson test with Type I error rate α



Fisherian *p*-value



Bayes factors

- NP & Fisher give H_0 a special role
- NP & Fisher optimize over θ, integrate over $D_{\rm hyp}$
- Bayes considers rival H_i symmetrically
- Bayes integrates over θ, uses only $D_{\rm obs}$

Bayes factors can only compare rival models;
they don't measure "goodness-of-fit"

Posterior predictive p-values are a BDA alternative
for measuring "suprisingness" of data for model checking;
they integrate over both data and parameter spaces

See "p-value note" online at 2016 CASt summer school site

Model averaging

Problem statement

$I = (M_1 \vee M_2 \vee \ldots)$ — Specify a set of models
Models all share a set of "interesting" parameters, ϕ
Each has different set of nuisance parameters η_i (or different prior info about them)
H_i = statements about ϕ

Model averaging

Calculate posterior PDF for ϕ:

$$\begin{aligned} p(\phi|D,\mathcal{C}) &= \sum_i p(M_i|D,\mathcal{C})\, p(\phi|D,M_i) \\ &\propto \sum_i \mathcal{L}(M_i) \int d\eta_i\, p(\phi,\eta_i|D,M_i) \end{aligned}$$

The model choice is a (discrete) nuisance parameter here

Useful for handling systematic error in estimation & prediction

Theme: Parameter Space Volume

Bayesian calculations sum/integrate over parameter/hypothesis space!

(Frequentist calculations average over *sample* space & typically *optimize* over parameter space.)

- Credible regions integrate over parameter space
- Marginalization weights the profile likelihood by a volume factor for the nuisance parameters
- Model likelihoods have Ockham factors resulting from parameter space volume factors

Many virtues of Bayesian methods can be attributed to this accounting for the "size" of parameter space. This idea does not arise naturally in frequentist statistics (but it can be added "by hand").

Roles of the prior

Prior has two roles

- Incorporate any relevant prior information
- Convert likelihood from “intensity” to “measure”
  → account for *size of parameter space*

Physical analogy

$$\begin{aligned} \text{Heat} \quad Q &= \int d\mathrm{r}\; c_v(\mathrm{r})\,T(\mathrm{r}) \\ \text{Probability} \quad P &\propto \int d\theta\; p(\theta)\mathcal{L}(\theta) \end{aligned}$$

Maximum likelihood focuses on the “hottest” parameters.
Bayes focuses on the parameters with the most “heat.”

A high-T region may contain little heat if its c_v is low or if its volume is small.

A high-$\mathcal{L}$ region may contain little probability if its prior is low or if its volume is small.

Supplement:

- Assigning priors
- Rule-based “objective” priors: Jeffreys, reference

Agenda

1. **Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$**
   - Confidence intervals vs. credible intervals
2. **Probability theory as generalized logic**
3. **Probability theory for data analysis: Three theorems**
4. **Inference with parametric models**
   - Parameter Estimation
   - Model Uncertainty
5. **Quick-looks**
   - Curve fitting & least squares
   - Measurement error & hierarchical/graphical models
   - Bayesian computation menu

Bayesian Curve Fitting & Least Squares

Setup

Data $D = \{d_i\}$ are measurements of an underlying function $f(x;\theta)$ at N sample points $\{x_i\}$. Let $f_i(\theta) \equiv f(x_i;\theta)$:

$$d_i = f_i(\theta) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma_i^2)$$

We seek to learn θ, or to compare different functional forms (model choice, M)

Likelihood

$$\begin{aligned} p(D|\theta, M) &= \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{d_i - f_i(\theta)}{\sigma_i}\right)^2\right] \\ &\propto \exp\left[-\frac{1}{2}\sum_i \left(\frac{d_i - f_i(\theta)}{\sigma_i}\right)^2\right] \\ &= \exp\left[-\frac{\chi^2(\theta)}{2}\right] \end{aligned}$$

Posterior

For prior density $\pi(\theta)$,

$$p(\theta|D,M) \propto \pi(\theta)\exp\left[-\frac{\chi^2(\theta)}{2}\right]$$

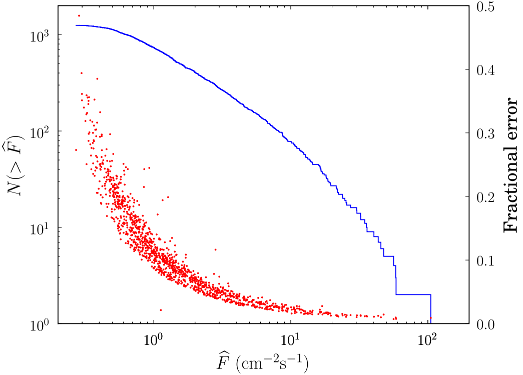
If you have a least-squares or χ^2 code:

- Treat $\chi^2(\theta)$ as $-2\log\mathcal{L}(\theta)$
- Bayesian inference amounts to exploration and *numerical integration* (by quadrature or Monte Carlo) of $\pi(\theta)e^{-\chi^2(\theta)/2}$

Agenda

1. **Motivating example: $\bar{x} \pm \sigma/\sqrt{N}$**
   - Confidence intervals vs. credible intervals
2. **Probability theory as generalized logic**
3. **Probability theory for data analysis: Three theorems**
4. **Inference with parametric models**
   - Parameter Estimation
   - Model Uncertainty
5. **Quick-looks**
   - Curve fitting & least squares
   - Measurement error & hierarchical/graphical models
   - Bayesian computation menu

Motivation: Complications with survey data

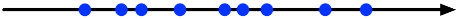


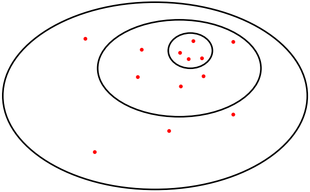
- *Selection effects* (truncation, censoring) — *obvious* (usually)
  Typically treated by "correcting" data
  Most sophisticated: product-limit estimators

- *"Scatter" effects* (measurement error, etc.) — *insidious*
  Typically ignored (average out? *No*—Eddington bias!)

Accounting for measurement error

Suppose $f(x|\theta)$ is a distribution for an observable, x (scalar or vector, $\vec{x} = (x, y, \ldots)$); and θ is unknown





From N precisely measured samples, $\{x_i\}$, we can infer θ from

$$\mathcal{L}(\theta) \equiv p(\{x_i\}|\theta) = \prod_i f(x_i|\theta)$$

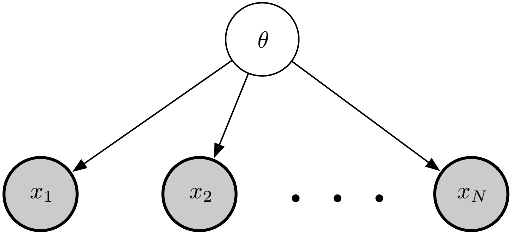
(A *binomial point process*)

$$p(\theta|\{x_i\}) \;\propto\; p(\theta)\mathcal{L}(\theta) \;=\; p(\theta, \{x_i\})$$

Posterior $\propto$ joint for params & data

Graphical representation

- Nodes/vertices = uncertain quantities (gray → known)
- Edges specify conditional dependence
- Absence of an edge denotes *conditional independence*

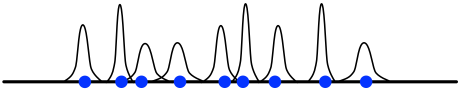


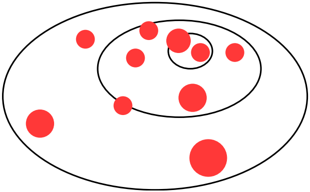
Graph specifies the form of the *joint distribution*:

$$p(\theta, \{x_i\}) \;=\; p(\theta)\, p(\{x_i\}|\theta) \;=\; p(\theta) \prod_i f(x_i|\theta)$$

Posterior from BT: $p(\theta|\{x_i\}) = p(\theta, \{x_i\})/p(\{x_i\})$

But what if the x data are *noisy*, $D_i = \{x_i + \epsilon_i\}$?





$\{x_i\}$ are now *uncertain (latent/hidden/incidental) parameters*

Member/item likelihoods quantify uncertainties: $\ell_i(x_i) = p(D_i|x_i)$

The joint PDF for *everything* is

$$\begin{aligned} p(\theta, \{x_i\}, \{D_i\}) &= p(\theta)\; p(\{x_i\}|\theta)\; p(\{D_i\}|\{x_i\}) \\ &= p(\theta) \prod_i f(x_i|\theta)\; \ell_i(x_i) \end{aligned}$$

The conditional (posterior) PDF for the unknowns is

$$p(\theta, \{x_i\}|\{D_i\}) = \frac{p(\theta, \{x_i\}, \{D_i\})}{p(\{D_i\})} \propto p(\theta, \{x_i\}, \{D_i\})$$

$$
\begin{aligned}
p(\theta, \{x_i\} | \{D_i\}) &\propto p(\theta, \{x_i\}, \{D_i\}) \\
&= p(\theta) \prod_i f(x_i|\theta)\, \ell_i(x_i)
\end{aligned}
$$

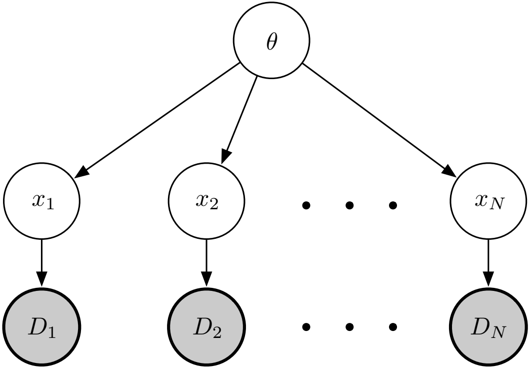
Marginalize over $\{x_i\}$ to summarize inferences for θ

Marginalize over θ to summarize inferences for $\{x_i\}$

Key point: *Maximizing over* x_i *(i.e., just using best-fit/MLE* $\hat{x}_i$*) and integrating over* x_i *can give very different results!*

(See Loredo (2004) for tutorial examples)

Graphical representation



$$
\begin{aligned}
p(\theta, \{x_i\}, \{D_i\}) &= p(\theta)\; p(\{x_i\}|\theta)\; p(\{D_i\}|\{x_i\}) \\
&= p(\theta) \prod_i f(x_i|\theta)\; p(D_i|x_i) \;=\; p(\theta) \prod_i f(x_i|\theta)\; \ell_i(x_i)
\end{aligned}
$$

A two-level *multi-level model* (MLM)

Hierarchical Bayes/MLMs in astronomy

- CASt 2014 Supplement Session — Includes discussion of selection effects
- Angie Wolfgang's lectures at CASt 2018 Astroinformatics School, CASt 2016 Summer School
- Survey of MLMs in astronomy: Bayesian astrostatistics: A backward look to the future (TJL 2013)
- `CUDAHM` C++ GPU software (Szalai-Gindl, Budavari, Kelly, TL); see Astron. & Comp. paper
- Stan and PyStan at MC-Stan.org
  Also see TL's StanFitter, AAS231-CosmicPopulations

Agenda

1. **Motivating example:** $\bar{x} \pm \sigma/\sqrt{N}$
   Confidence intervals vs. credible intervals

2. **Probability theory as generalized logic**

3. **Probability theory for data analysis: Three theorems**

4. **Inference with parametric models**
   Parameter Estimation
   Model Uncertainty

5. **Quick-looks**
   Curve fitting & least squares
   Measurement error & hierarchical/graphical models
   Bayesian computation menu

Bayesian computation menu

Large sample size, N: Laplace approximation

- Approximate posterior as multivariate normal → det(covar) factors
- Uses ingredients available in χ^2/ML fitting software (MLE, Hessian)
- Often accurate to $O(1/N)$ (better than $O(1/\sqrt{N})$)

Modest-dimensional models ($m \lesssim 10$ to 20)

- Quadrature, cubature, adaptive cubature
- IID Monte Carlo integration (importance & stratified sampling, adaptive importance sampling, quasirandom MC)

High-dimensional models ($m \gtrsim 5$): Non-IID Monte Carlo

- Posterior sampling — create RNG that samples posterior
  - Markov Chain Monte Carlo (MCMC) is the most general framework
- Sequential Monte Carlo (SMC)
- Approximate(ly) Bayesian computation (ABC)
- . . .

Recap of key ideas

Probability as generalized logic

Probability quantifies the *strength of arguments*

To appraise hypotheses, calculate probabilities for arguments from data and modeling assumptions to each hypothesis

Use *all* of probability theory for this

Bayes's theorem

$$p(\text{Hypothesis} \mid \text{Data}) \propto p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})$$

Data *change* the support for a hypothesis ∝ ability of hypothesis to *predict* the data

Law of total probability

$$p(\text{Hypothes}\underline{\textbf{es}} \mid \text{Data}) = \sum p(\text{Hypothes}\underline{\textbf{is}} \mid \text{Data})$$

The support for a *compound/composite* hypothesis must account for all the ways it could be true